Named Entity Recognition and Classification for Gujarati Language

Komil Vora
Information Technology Department
V.V.P Engineering College
Rajkot-Gujarat-India
Email: komil.vora@vvpedulink.ac.in

Dr. Avani Vasant
Computer Science & Engineering Department
Babaria Institute of Technology & Science
Vadodara-Gujarat-India
Email: avani.vasant@gmail.com

Rachit Adhvaryu
Information Technology Department
V.V.P Engineering College
Rajkot-Gujarat-India
Email: rvadhvaryu@gmail.com

***Abstract*—Named Entity Recognition (NER) is a method to search for a particular Named Entity (NE)[1] from a file or an image, recognize it and classify it into specified Entity Classes like Name, Location, Organization, Numbers and Others Categories. It is the most useful element of the technique known as Natural Language Processing (NLP) which makes text extraction very easy [2]. In this paper, we focus on using Hidden Markov Model (HMM) based techniques to recognize the Named Entity (NE) for Gujarati language. The main aim of using HMM is that it provides better performance and can be easily implemented for any languages. A remarkable amount of work has been carried out for many languages like English, Greek, Chinese etc. But, still a wide scope is open for Indian Origin Languages like Hindi, Gujarati, Devanagari etc. As Gujarati is not only the Indian Language, but a language that is most spoken in Gujarat. Thus, in this paper, we emphasis on proposing a NER based scheme for Gujarati Language using HMM.**

Keywords. Named Entity Recognition (NER), Hidden Markov Models (HMM), Gujarati Language

I. Introduction

Named entity recognition (NER) is the process of recognizing entity (e.g. Person, Place, Organization etc.) associated with a particular word of a document, it also classify word in different categories. NER has a significant part in natural language processing (NLP) applications like information retrieval, machine learning, survey systems, automatic summarization etc.[3]

NER systems uses language grammar-based techniques as well as arithmetical models (i.e. machine learning). Handcrafted grammar-based systems usually obtain enhanced accuracy. Arithmetical NER systems usually require a very large manually interpreted trained data. Named Entity Recognition (NER) is used to organize every word in a document into predefined categories [4]. It falls under the category of ̈data mining, which mines specific types of information or data from documents. Since entity names are the major part of the document, NER is a very key step towards more intellectual data mining and its management.

A ample amount of research work has been carried out for NER in Various Languages like English, Greek, Chinese etc. with very little variations in text size, shape and other dimensions [5]. India is only the country with lots of Languages with very wide variety of text characters and shapes. Implementing NER for Indian origin languages is challenging task. We are focusing on implementing NER for Gujarati Language using HMM.

II. Methods and Related Work for NER

[6] There are usually two methodologies for Named Entity Recognition which includes:

1) Rule Based Approach
2) Machine Learning Based Approach

1) **Rule Based Approach**

Rule based approaches are the most efficient for NER systems. It generally uses rules written manually by the experts. It provides high accuracy as compared to other approaches. Basically Rule based approach is classified into following:

a) Linguistic Approach
b) List Look-up Approach

The main drawback of Rule Based Approaches are:[6]

- A vast expertize is required for generating rules for a particular domain.
- An implementation of a system requires very much effort and time.
- The rules defined for one language or domain cannot be transferred or used for other language or domain.
- Implementing a small modification is very complex.

2) **Machine Learning Based Approaches**

Machine Learning based approaches are fully dependent on arithmetical models to predict Named Entity (NE) in a given document. A great amount of Metadata information is required to make this approach successful and worthy. There are mainly 5 machine learning approaches:

a) Hidden Markov Model
A hidden Markov model (HMM) is an algebraic Markov model where it is assumed that the system is a developed with overlooked states. A HMM can be presented as the simplest dynamic Bayesian network [8]. A Hidden Markov Model (HMM) can be

978-1-5090-2029-4/16/$31.00 @2016 IEEE

considered as an outline of a fusion model where the overlooked variables which controls the combination component that is taken for observation are through markov process rather independent.

Hidden Markov models are mostly known for their applications such as speech & handwriting recognition, movement recognition, part of-speech labeling, musical score following, bioinformatics etc.[8]

HMM is proficient of allocating semantic labels to tokens over a inputs; this is beneficial for text-related tasks that involve some ambiguity, including part-of-speech labeling, text separation, named entity recognition (NER) and data mining tasks[7]. However, most of the natural language processing tasks are reliant on determining an entity associated with information. An example would be word "Rajkot" is a Location; therefore word "Rajkot" will receive a tag of Location.

b) Maximum Entropy

Maximum Entropy (ME) is also known as provisional probability based sequence model. In this model, many features are mined from the words and the dependencies of word on other word is maintained [9].

In Maximum Entropy model, we consider least biased known facts that maximize the entropy [9]. Every source takes an assessment feature as an input and spreading over the following state as an output. The subsequent outputs are related with the states.

The possibility adaptation leaving any given state must be equal to one, so, it has an effect towards that states with very few outward evolutions. The state with one outward state evolution will deny every annotations [9].

c) Conditional Random Fields

CRF is a kind of particular probability model. CRFs are directionless pictorial models and also known as unplanned (random) fields which are used to estimate the provisional possibility of values on allotted outward nodes given the values allotted to other allotted input nodes [10]. It is used to translate known relationships between readings and create reliable analysis.

Conditional random fields (CRFs) are a class of arithmetical methods often useful in pattern discovery and machine learning, where they are used for planned estimations. A normal classifier estimates a label for a one sample regardless of "adjacent" samples. A CRF can be useful in this context; e.g., the linear chain CRF generally in natural language processing estimated the label sequences for respective input samples. In computer vision, CRFs are regularly used for element identification and image seperation.

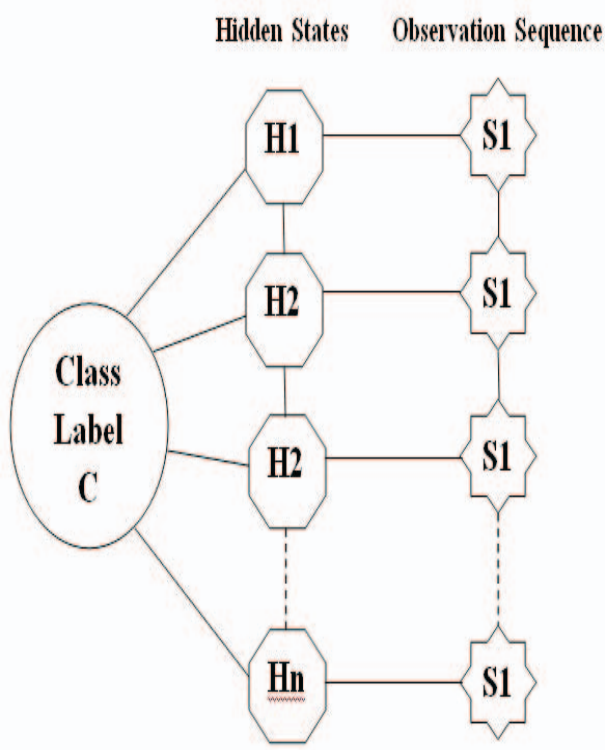

Fig. 1: Conditional Random Field

d) Support Vector Machine

The SVM is a unique approach which is used for optimistic and pessimistic instances to gain knowledge of differences among the two classes. The SVMs are also known to vigorously manage huge feature sets and to create models that exploit their general forms. SVM is a scheme for proficiently training linear machines in kernel-based feature spaces, while regarding the visions of simple concepts and maximizing optimization concepts [10].

e) Decision Trees

Decision Trees is the most widely used and effective technique for categorization and forecasting. Rules are used for artificial intelligence and neural network which can be easily understood by the humans and can be directly used with database languages.

It is a classification method in the form of tree where each node of the method is represented by a leaf, expressing value of outward attribute, a conclusion, expressing the text which is to be carried on one branch and a sub tree, expressing the probable results of the text. It is a very unique method to obtain knowledge on classification [10].

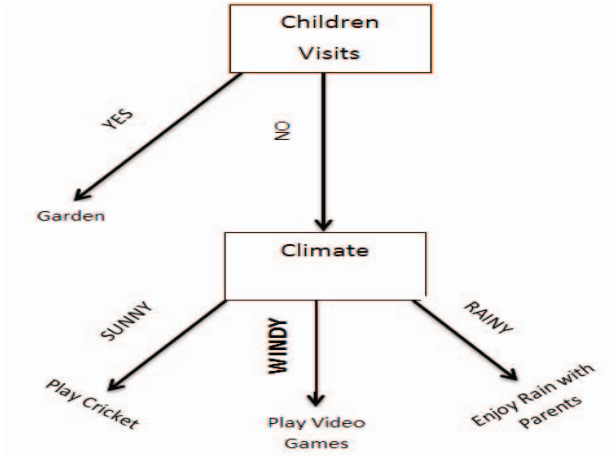

Fig. 2: Decision Trees

III. Indian Languages

India is the only country which has many different languages. Each language has specific characteristic which differentiates it from the other languages. There is a large variation in each language. The most common and widely spoken languages not only in India, but across the world are Hindi, Tamil, Punjabi, Bengali and Gujarati [5].

A. Challenges

Implementing NER in Indian Languages is different and more difficult than implementing it in Roman Script Languages. Following are the reasons for it[4]:

1) There is no concept of Capitalization. All words are written in similar case.
2) There is uncertainty and no particular standard has been defined to write spellings or words.
3) Indian languages are having no specific order of words.
4) Indian languages are considered to be poor languages as a good dictionary, web source and well-built language analyzer is not available yet.
5) Still a lot of technological enhancement is required in Indian languages.

B. Current Scenario

While implementing NER in Indian Languages, the problems that are faced are:

1) NER for one language cannot be implemented for the other language.
2) implementing same NER for different languages requires too much work and efforts.
3) In some cases, Rule based Approach gives optimal solution with high accuracy. But, Language experts are required to generate rules for each language.
4) As Indian Languages can be written in any order, new words can be easily created and thus maintaining the word list is a big task. But still a Gazetter method gives an acceptable output.
5) Gazetter method takes a lot of time to search and recognize the Named Entity (NE)[4].

As discussed earlier, we will work on Gujarati Language and implement NER using Hidden Markov Model (HMM) because:

1) HMM is not a particular language or domain based method. It can be used for any language.
2) The NER system developed using HMM can be easily implemented and understood.
3) Language experts are not required. A person with very little knowledge of a language can easily implement NER system.
4) HMM is dynamic in nature. It can be used in any manner or as per requirement [6].

IV. Our Proposed Approach

Our approach mainly focuses on Gujarati Language. Gujarati Language is very large in terms of entities or words. Each character in Gujarati language is different than other character in terms of shapes, sizes and other dimensions. Our main aim is to recognize gujarati named entity (NE) from the group of words and classify each word based on the different classes like Name, Location, Others etc. For this, we will implement Named Entity Recognition (NER) using Hidden Markov Model (HMM).

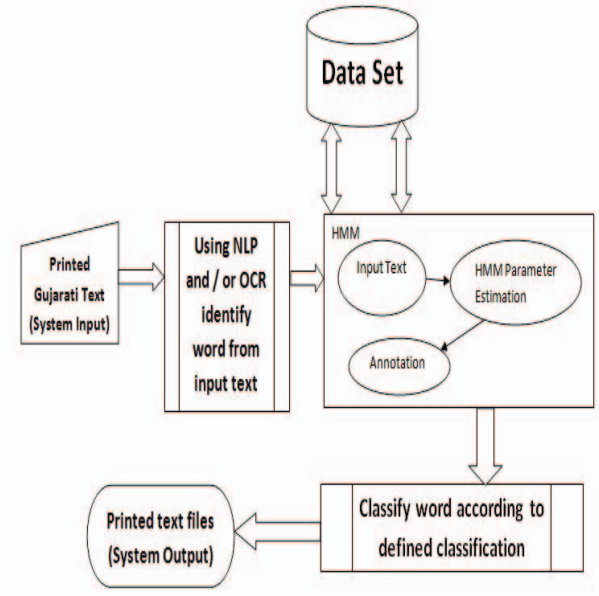

Fig. 3: Steps of Recognition and Classification

Algorithm

With reference to above image our proposed system will work as per the following steps:

1) We will prepare a dataset with different set of classes (e.g.: Name, City, Number etc.).
2) System will take image of printed Guajarati text as a system input.

3) Using Natural Language Processing and / or optical character recognition, system will separate all the words from an image and generate a text file.
4) This text file will be given as an input to HMM.
5) Using HMM Algorithm, the tags associated with the words will be identified.
6) Once the tags are identified, system will compare the words associated with the identified tags with the defined dataset.
7) If the match is found, system will classify those words based on the defined set of classes.
8) Final output of the system will be printed text files with all the possible classification of identified words from the input image.

Printed Gujarati Text	Name of Entity
મોહનદાસ કરમચંદ ગાંધી	Name
રાજકોટ	Location
મેડીકલ કાઉન્સીલ ઓફ ઇન્ડિયા	Organization
૩૨, એકસો એક (૧૦૧)	Numbers
તા. ૧/૧/૨૦૦૦ (૧ જાન્યુઆરી)	Others

Fig. 4: Recognized Named Entity

Fig. 4 depicts our proposed system. As shown in the algorithm steps, system is expected to work on Gujarati Text with HMM and Named Entity will be recognized. Related references can be found in [11]; where work done on other Indian Languages like Hindi, Urdu and Marathi.

V. Conclusion and Future Work

We proposed an approach for NER in Gujarati Language using HMM. The approach is unique and is believed to give optimum output after implementation.

The mentioned approach works well on printed Gujarati words. In case of hand written Gujarati characters, it may or may not work efficiently as the writing technique varies from person to person. A wide variation in the text size, shape and dimensions of the text is found. Thus, in future, a system can be proposed to expand the current work for hand written Gujarati characters which provides an optimum output in any cases.

References

[1] Sudha Morwal and Nusrat Jahan, Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages, in International Journal of Advanced Research in Computer Science and Software Engineering, Vol-3,Issue-4,pp. 971-975
[2] Sudha Morwal, Nusrat Jahan and Deepti Chopra, Named Entity Recognition using Hidden Markov Model (HMM), in International Journal on Natural Language Computing(IJNLC), December 2012, pp. 15-23.
[3] Feifan Liu, Jun Zhao, Bibo Lv, Bo Xu and Hao Yu, Product Named Entity Recognition Based on Hierarchical Hidden Markov Model.
[4] GuoDong Zhou and Jian Su, Named Entity Recognition using an HMM-based Chunk Tagger, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 473-480.
[5] Lin-Yi Chou, Techniques to incorporate the benefits of a Hierarchy in a modified hidden Markov model, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, July 2006, pp 120-127.
[6] Sujan Kumar Saha Sanjay Chatterji Sandipan Dandapat. "A Hybrid Approach for Named Entity Recognition in Indian Languages".
[7] Darvinder kaur, Vishal Gupta. "A survey of Named Entity Recognition in English and other Indian Languages", International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.
[8] Alireza Mansouri, Lilly Suriani Affendey and Ali Mamat, Named Entity Recognition Approaches, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008, pp. 339-344
[9] Pramod Kumar Gupta, Sunita Arora "An Approach for Named Entity Recognition System for Hindi: An Experimental Study" in Proceedings of ASCNT - 2009, CDAC, Noida, India, pp. 103 - 108.
[10] Daljit Kaur and Ashish Verma, Survey on Name Entity Recognition Used Machine Learning Algorithm, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5875-5879
[11] Sudha Morwal and Nusrat Jahan, Named Entity Recognition Using Hidden Markov Model(HMM): An Experimental Result on Hindi, Urdu and Marathi Languages , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 4, April 2013, 671-675